Use of semantics in large scale vertical search

April 23rd, 2008

Per Gunnar Auran, PhD,

Yahoo! Technologies Norway AS (YTN)

Outline

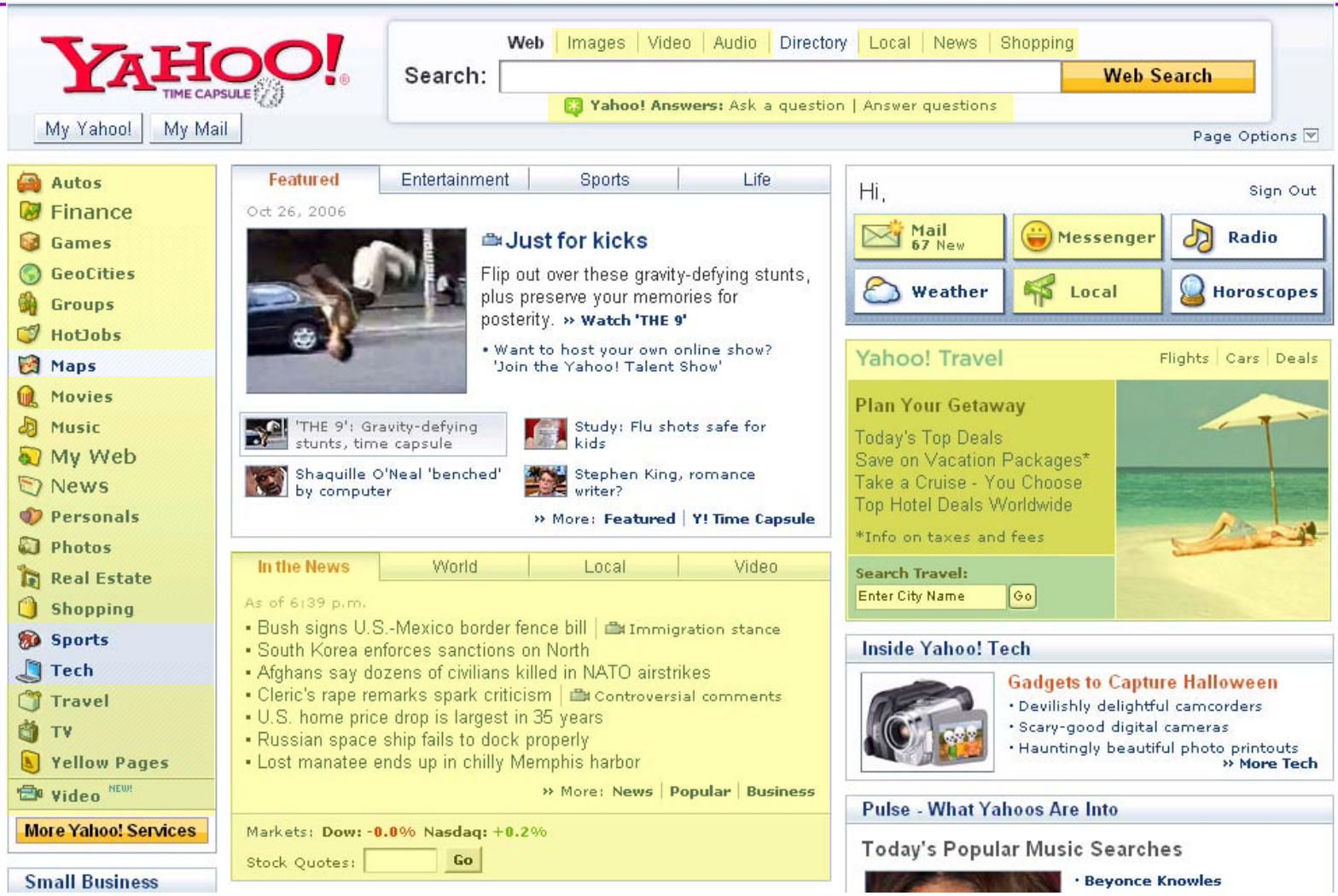
- **Introduction** (10 min)
  - Yahoo! Technologies Norway: Background
  - Search platform overview
  - Verticals vs. web search

- Semantics & large scale search (15-20 min)
  - *Simplistic approach to semantic search that scales*
  - Examples:
    - personalized search, local search, *shopping search*

Introduction: Yahoo! Technologies Norway AS

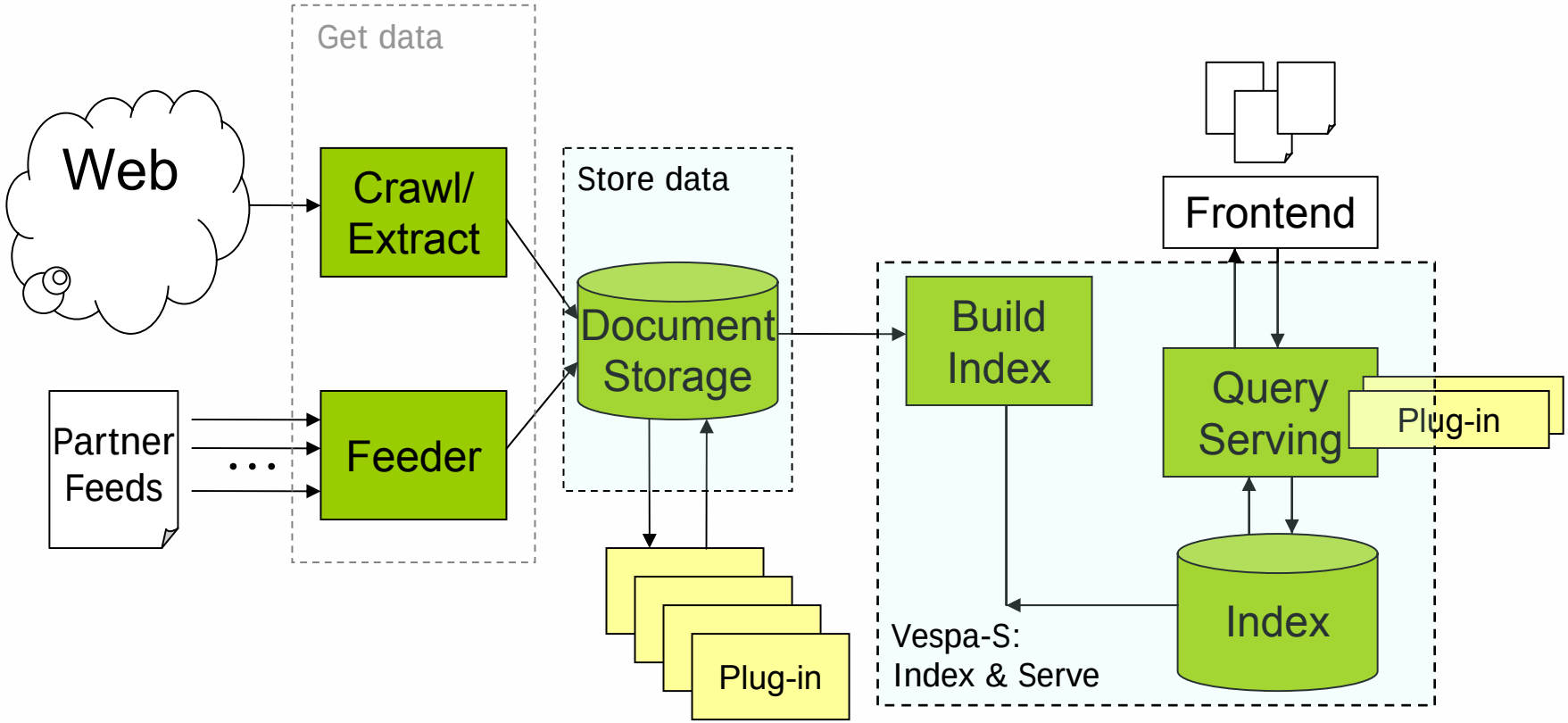
- 1997-2003 FAST Search & Transfer ASA
  - FTPSearch => AllTheWeb
  - Web Search: 40+ people
    - Development in TRD
    - Operations in US
- 2003: “Year of Consolidation”
  - April: Acquired by Overture, the Internet Marketing Leader
  - October: Overture acquired by Yahoo!
  - Goal: Make a vertical search platform (Vespa) for Yahoo!
- 2007: ***Vespa de facto std search engine for Y! verticals world wide***
  - 100+ installations

Introduction: What we do in Trondheim?



Vespa search platform overview

End-to-end Solution



Search application configuration

Search definitions => XML search schema

- The document type(s)
  - *How to index it?*
    - Fields, stemming etc.
  - *How to search it?*
    - Default vs. special index
  - *What to present?*
    - Data in result set

```
# A simple example
search book {

  document book {
    field title type string {
      indexing:   index | summary
      rank-type:  identity
      rank-boost: 1000
    }

    field price type int {
      indexing: summary | attribute
    }
  }

}
```

Some numbers: Huge in traffic vs. data

- Yahoo! Answers:
  - **2700 queries per second** (QPS)
  - 45 millions queries/answers
- Yahoo! Mail:
  - **30.000 emails per second**
  - 200 million users, 200+ billions email searchable
  - Hundreds of terabytes data
  - Lower QPS rate (few hundreds)

Vertical vs. web search

Verticals: Specialized web services based on search

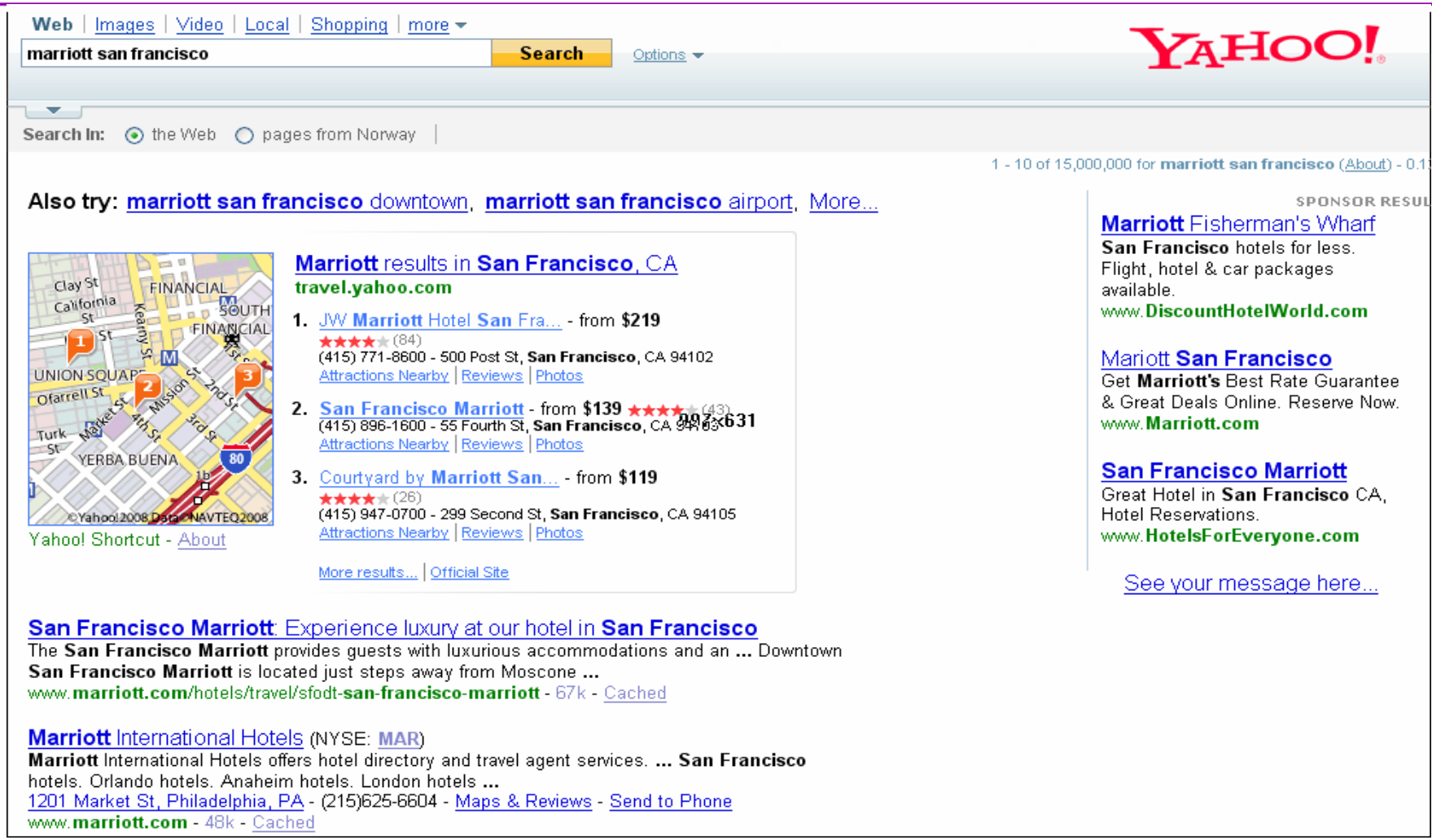
	Vertical search	Web search
Index size	Smaller and specialized	Global and general
Document type	Typically more structured /DB legacy	Typically less structured
Relevance	*Highly customizable* Relevance enhanced by – Constrained context + structure	*Relatively fixed algorithm* & document model Popularity-based
Comprehensiveness	Focused/deeper crawling + feeds	Broad/surface crawling
Freshness	Customizable schedules From seconds to months	Fixed schedule Days on average
Presentation	Structured, Navigational –Sorting & grouping –Clustering & collapsing	Flat list

Databases ***matches precisely*** Search ***scales well***

Database search => Web Search => Verticals = Integration of search & DB technology

Convergence to one technology: ***HYBRID SEARCH***

Blending vertical & web search results



Vertical relevance = f(domain)

Data Structure + Rank Dimensions:

- **Q**uery term match quality:
  - Field, position, statistics, proximity, term overlap quality
- **T**ime dimension
  - Document freshness, refresh rate, temporal/seasonal effects
- **L**ocation/Distance
  - Geographical, virtual (server/host), topical (category/ontology)
- **A**ttributes
  - Filter, boost, aggregate, sort
- **D**ocument quality
  - Authority, popularity, information value, layout/typesetting

Outline

- Introduction
  - Yahoo! Technologies Norway: Background
  - Search platform overview
  - Verticals vs. web search

- **Semantics & large scale search**
  - *Simplistic approach to semantic search that scales*
  - Examples:
    - personalized search, local search, *shopping search*

Semantics & large scale search

Bag of keywords vs. structure ...

– ”sony digital cameras”

=> brand:sony item:”digital camera”

– ”jobs in bangalore”

=> listing:job location:bangalore

– “from san francisco to paris on april 23rd”

=> depart:”san francisco” arrive:paris date:04-23-2008

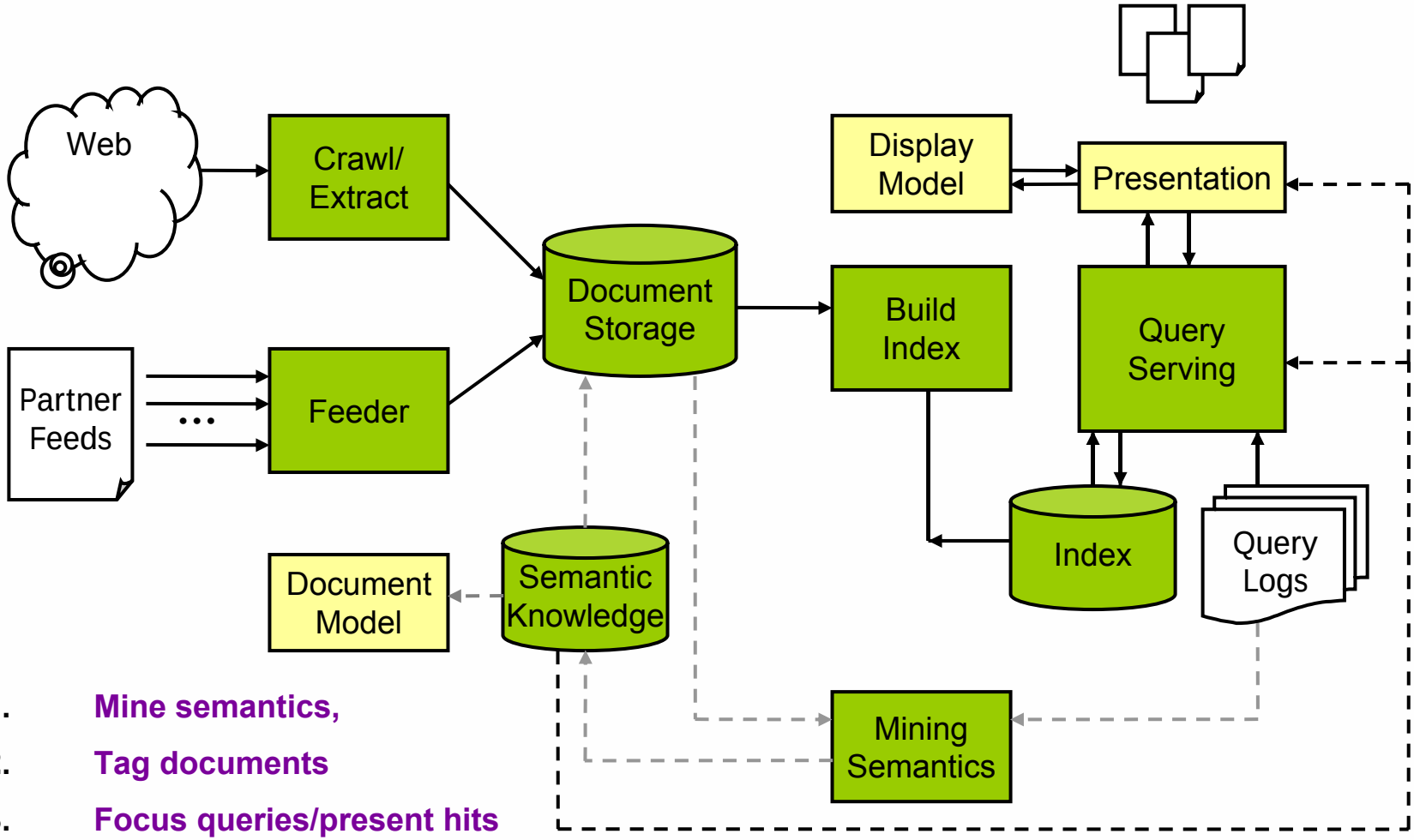
– ”restaurants in geary street”

⇒listing:restaurant address:”geary street”

Can we make our applications understand this?

- *When we can:* ***match meaning directly***
- *When we cannot:* ***default to keyword based matching***

Semantic search steps:



1. **Mine semantics,**
2. **Tag documents**
3. **Focus queries/present hits**

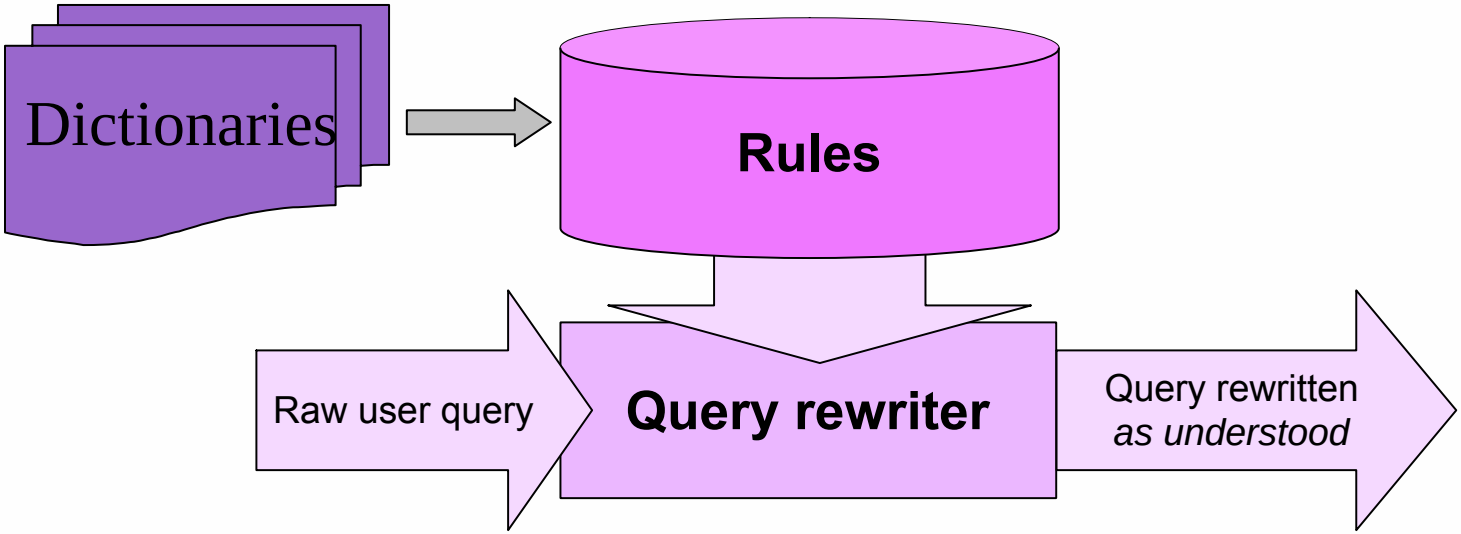
Vespa semantic search priorities

1. **Query rewrites**
   => Focused search / high precision
2. **Entity dictionary generation**
   => Toolbox for query log and feed mining
3. ***Flexible ranking model***
4. Content tagging:
   => Match precision & entity normalization
5. *Resultset processing:*
   => *Conditional ordering of hits*

Query rewriting support in Vespa

The rule based query rewriting language in Vespa-S

- Application developers represent the domain and linguistic knowledge using rules
- Queries are rewritten based on these rules



Writing semantic rules

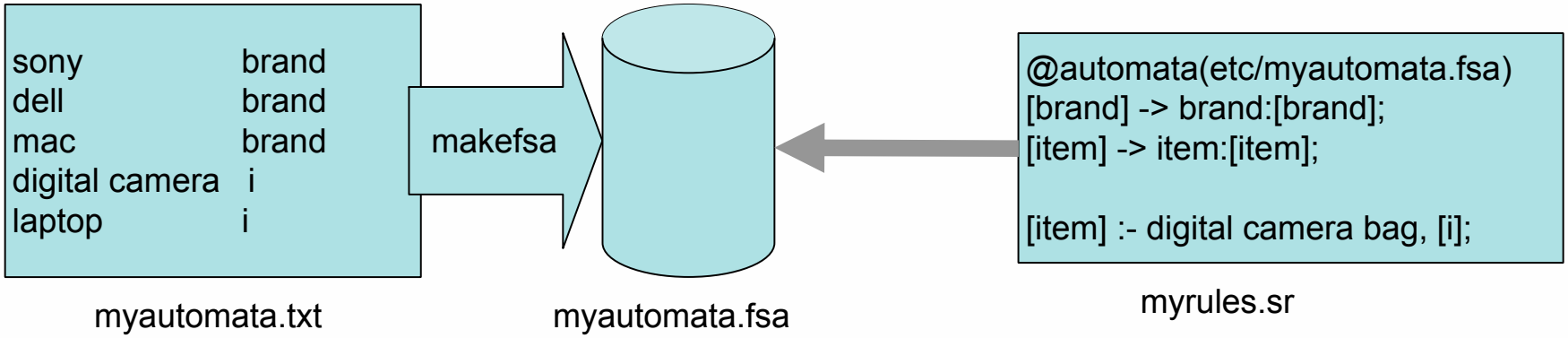
- Generic form: **`condition operator production;`**
- Conditions may refer to *named conditions*
- Productions may refer to what's matched in the condition

```
# + means add
lotr +> movienames:lotr
# - means replace
lotr -> lord of the rings;

# If we find a "car", replace it by carindex:"car"
[car] -> carindex:[car];
# Which phrases are a "car"
[car] :- audi, bmw, volvo;
```

Using dictionaries in rules

- Representing very many words is not done efficiently in rules
- Hence, the rule system uses a finite state automata library
  - Rules can use wordlists compiled into automatas



Semantic rule summary

- Semantic rules can:
  - **replace or add** terms
  - add **rank effects or required** matches
  - use **dictionaries** that support multiple **entities**
  - be **specified query time** (rulebases)
  - be used to
    - *choose document types & ranking*

Note: This is about efficient use of semantic/domain knowledge for scalable search, semantic analysis comes prior to this.

Semantics & large scale search (E1)

Example 1: Personalization / Context focused search

- ***Semantics:*** *Understanding content and relation to queries*
  - Term vector representation of documents & their topics
- ***Personalization****: Understanding individual user intent*
  - Individual preference: Queries and ”searchmarks”
- ***Personalized search:***
  - *Focusing search towards tagged content based on individual preference, targeting queries with multiple interpretations:*
    - jaguar => car / mac os x / cat / music / atari console

Query refinement & personalization

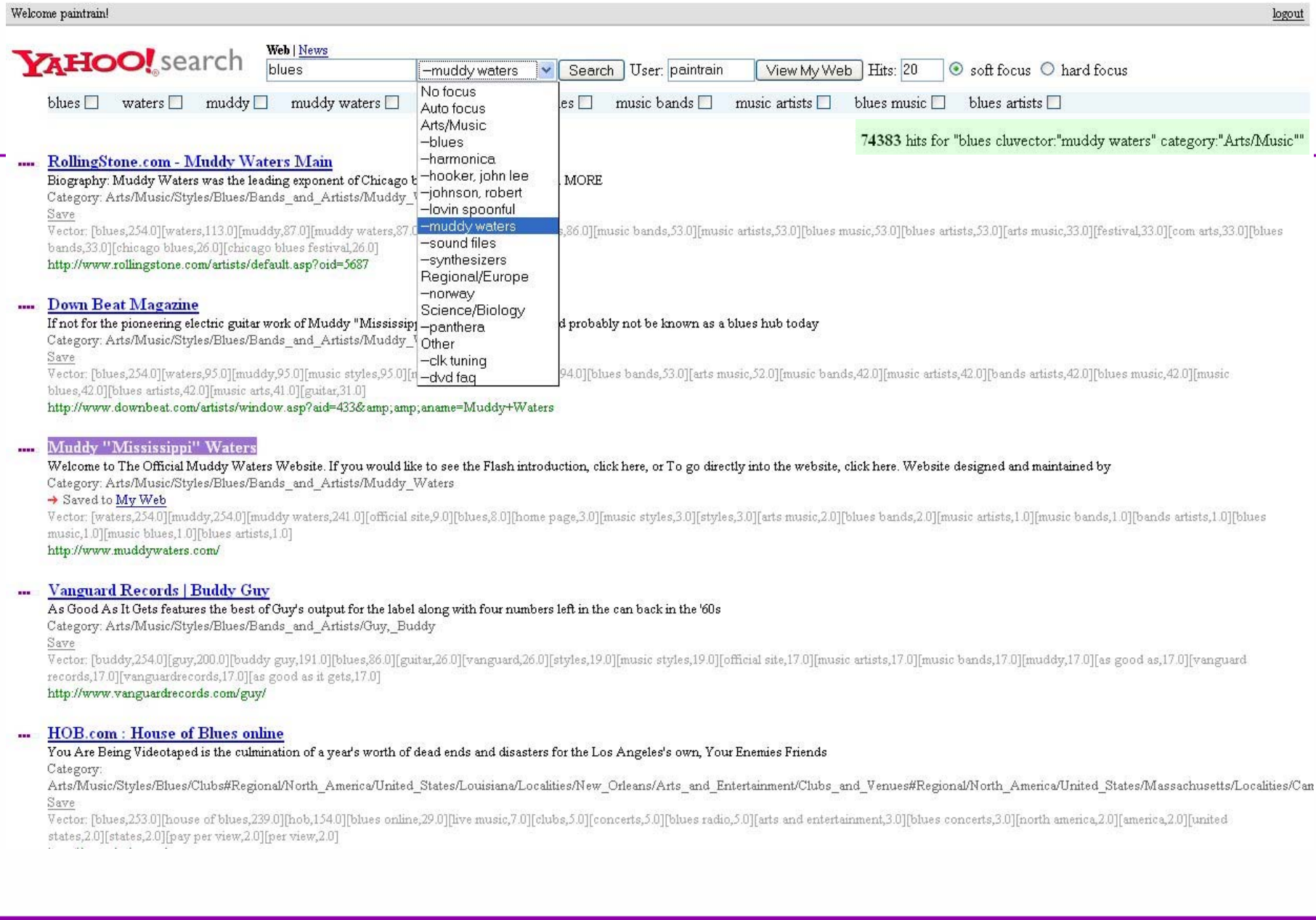
Query Q: Adding user preference focus

Q => Q + category:<topic unit>

+ interest:<doc vector unit>

- *Focus terms are added as optional terms with full rank effect :*
  - *improves precision, but recall is unchanged.*

This prototype led to the development of the generalized rule-based query rewrite engine...


Welcome paintrain!
logout
YAHOO! search
Web | News
blues
–muddy waters
Search
User: paintrain
View My Web
Hits: 20
soft focus
hard focus
No focus
Auto focus
Arts/Music
–blues
–harmonica
–hooker, john lee
–johnson, robert
–lovin spoonful
–muddy waters
–sound files
–synthesizers
Regional/Europe
–norway
Science/Biology
–panthera
Other
–clk tuning
–dvd faq
blues
waters
muddy
muddy waters
music bands
music artists
blues music
blues artists
74383 hits for "blues cluvector:"muddy waters" category:"Arts/Music""
RollingStone.com - Muddy Waters Main
Biography: Muddy Waters was the leading exponent of Chicago b
MORE
Category: Arts/Music/Styles/Blues/Bands_and_Artists/Muddy_
Save
Vector: [blues,254.0][waters,113.0][muddy,87.0][muddy waters,87.0
,86.0][music bands,53.0][music artists,53.0][blues music,53.0][blues artists,53.0][arts music,33.0][festival,33.0][com arts,33.0][blues bands,33.0][chicago blues,26.0][chicago blues festival,26.0]
http://www.rollingstone.com/artists/default.asp?oid=5687
Down Beat Magazine
If not for the pioneering electric guitar work of Muddy "Mississip
d probably not be known as a blues hub today
Category: Arts/Music/Styles/Blues/Bands_and_Artists/Muddy_
Save
Vector: [blues,254.0][waters,95.0][muddy,95.0][music styles,95.0][
94.0][blues bands,53.0][arts music,52.0][music bands,42.0][music artists,42.0][bands artists,42.0][blues music,42.0][music blues,42.0][blues artists,42.0][music arts,41.0][guitar,31.0]
http://www.downbeat.com/artists/window.asp?aid=433&aname=Muddy+Waters
Muddy "Mississippi" Waters
Welcome to The Official Muddy Waters Website. If you would like to see the Flash introduction, click here, or To go directly into the website, click here. Website designed and maintained by
Category: Arts/Music/Styles/Blues/Bands_and_Artists/Muddy_Waters
Saved to My Web
Vector: [waters,254.0][muddy,254.0][muddy waters,241.0][official site,9.0][blues,8.0][home page,3.0][music styles,3.0][styles,3.0][arts music,2.0][blues bands,2.0][music artists,1.0][music bands,1.0][bands artists,1.0][blues music,1.0][music blues,1.0][blues artists,1.0]
http://www.muddywaters.com/
Vanguard Records | Buddy Guy
As Good As It Gets features the best of Guy's output for the label along with four numbers left in the can back in the '60s
Category: Arts/Music/Styles/Blues/Bands_and_Artists/Guy,_Buddy
Save
Vector: [buddy,254.0][guy,200.0][buddy guy,191.0][blues,86.0][guitar,26.0][vanguard,26.0][styles,19.0][music styles,19.0][official site,17.0][music artists,17.0][music bands,17.0][muddy,17.0][as good as,17.0][vanguard records,17.0][vanguardrecords,17.0][as good as it gets,17.0]
http://www.vanguardrecords.com/guy/
HOB.com : House of Blues online
You Are Being Videotaped is the culmination of a year's worth of dead ends and disasters for the Los Angeles's own, Your Enemies Friends
Category:
Arts/Music/Styles/Blues/Clubs#Regional/North_America/United_States/Louisiana/Localities/New_Orleans/Arts_and_Entertainment/Clubs_and_Venues#Regional/North_America/United_States/Massachusetts/Localities/Can
Save
Vector: [blues,253.0][house of blues,239.0][hob,154.0][blues online,29.0][live music,7.0][clubs,5.0][concerts,5.0][blues radio,5.0][arts and entertainment,3.0][blues concerts,3.0][north america,2.0][america,2.0][united states,2.0][states,2.0][pay per view,2.0][per view,2.0]

Semantics & large scale search (E2)

Example 2: Local search

Basic entities

- **Business name:** specific business
- **Business category:** business/service type
- **Location:**
  location terms + geo-location for distance

Query rewrite examples

```
# Example semantic rules

@default
@automata(etc/vesparules.fsa)

# Recognize citystate first
[C] [S] -> $city:[C] $state:[S];

[B] +> $busname:[B];
[T] +> $bustopic:[T];
[C] +> $buscity:[C];
```

```
pizza riverside            => [RANK (AND pizza riverside) busname:riverside bustopic:pizza buscity:riverside]
pizza riverside ca         => [RANK pizza state:ca city:riverside bustopic:pizza]
best sushi santa clara ca  => [RANK (AND best sushi) state:ca city:santa°clara busname:sushi bustopic:sushi]
movie theatre santa clara ca => [RANK (AND movie theatre) state:ca city:santa°clara bustopic:"movie theatre"]
pizza chicago san jose ca  => [RANK (AND pizza chicago) state:ca city:san°jose busname:chicago
                               bustopic:pizza bustopic:chicago buscity:chicago]
new york pizza chicago il  => [RANK (AND new york pizza) state:il city:chicago busname:"new york"
                               bustopic:"new york" bustopic:pizza buscity:"new york"]
pizza chicago new york ny  => [RANK (AND pizza chicago) state:ny city:new°york busname:chicago
                               bustopic:pizza bustopic:chicago buscity:chicago]
```

Semantics & large scale search (E3)

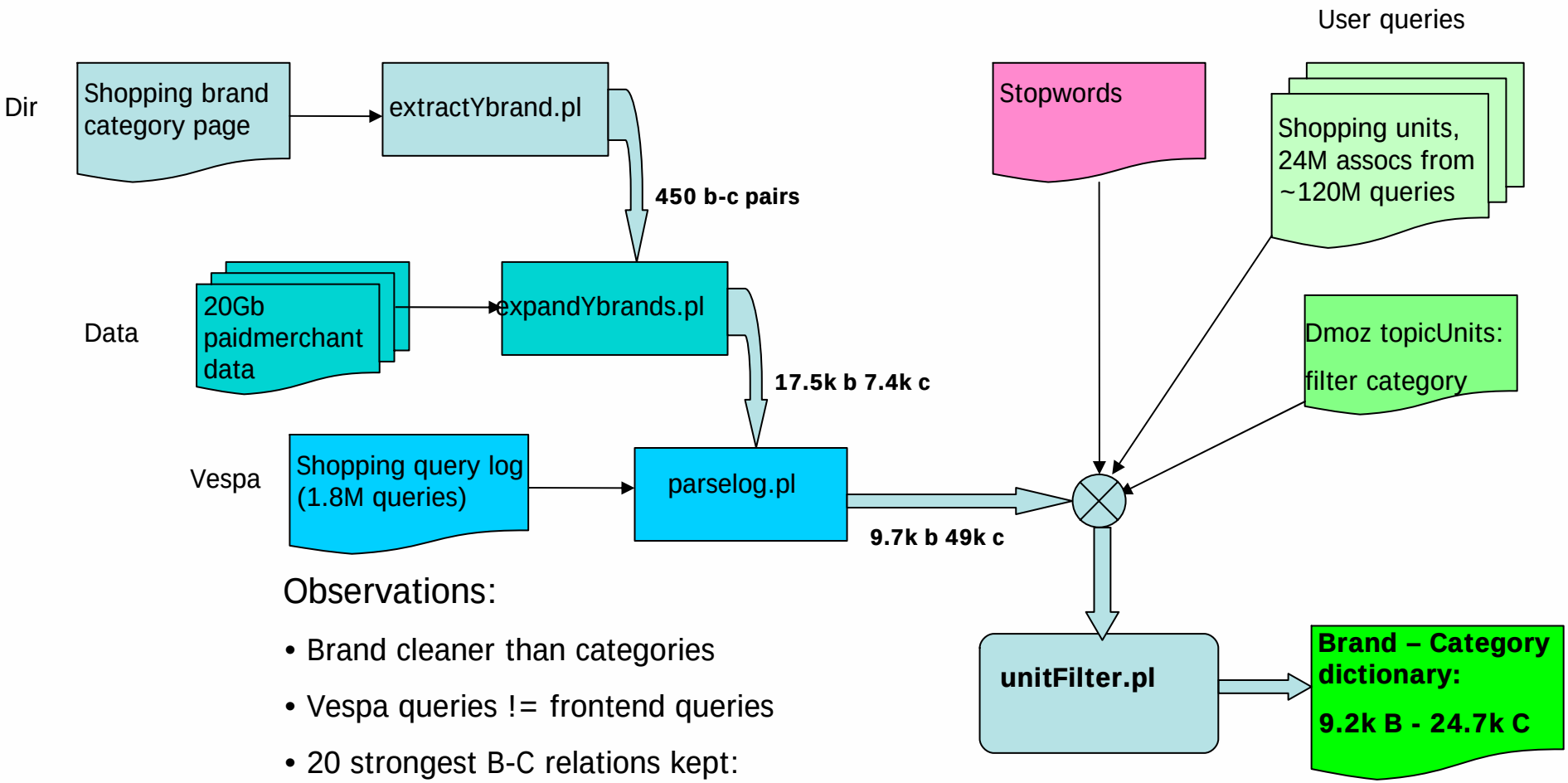
Example 3: Shopping / Relevant sorting

– *The primary ordering criterion != relevance, but results should be relevant*

Components

1. Query parser / rewriter
    - Simple query logic: brand/category match
2. Document tagger:
    - Entity normalization of ***brands and categories***
3. Brand and category dictionary
    – ***same one*** needed for both previous tasks

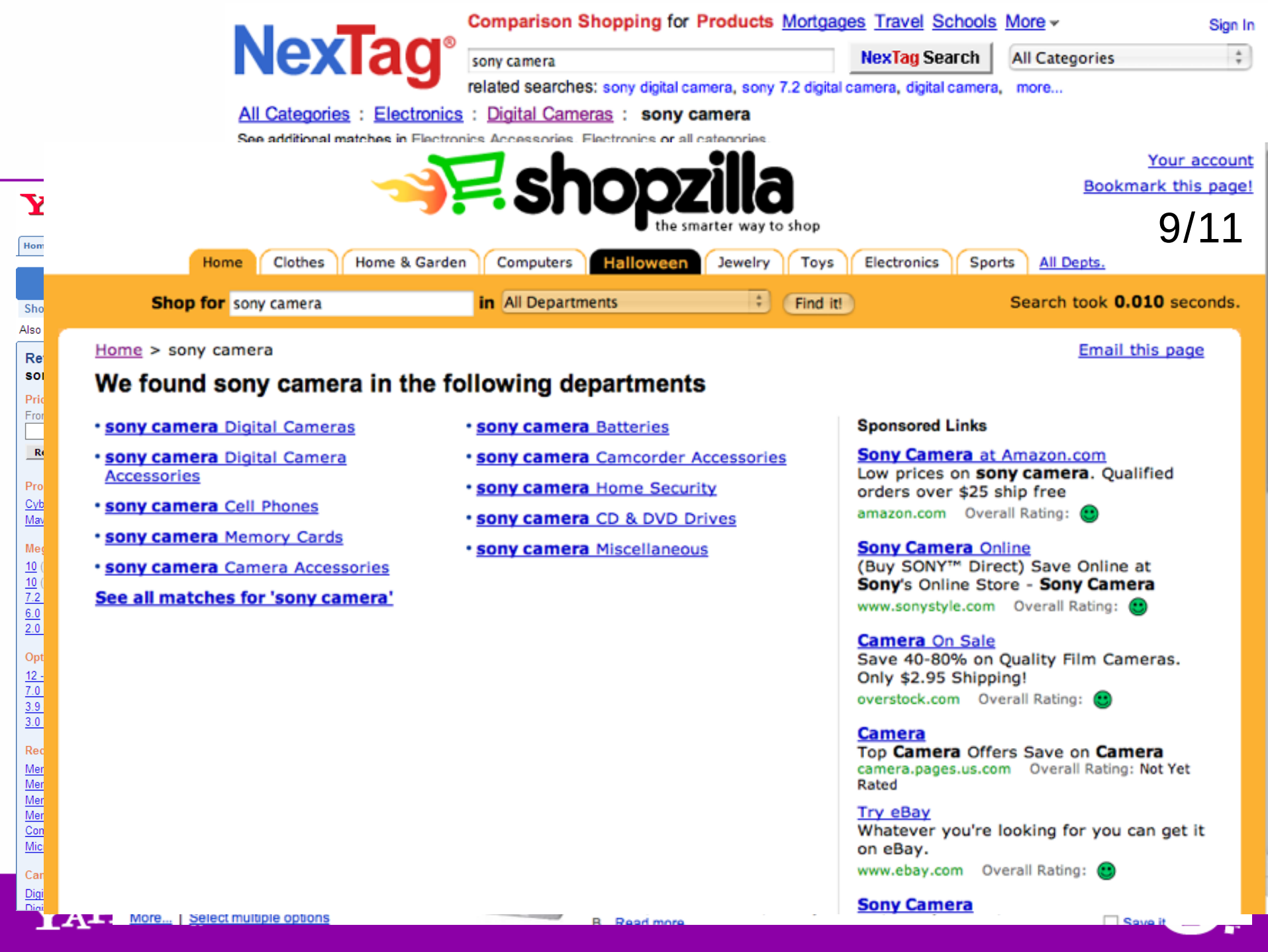
Shopping: Making an entity dictionary



Observations:

- Brand cleaner than categories
- Vespa queries != frontend queries
- 20 strongest B-C relations kept:

sony\tC;camera;dvd;lcd;digital;accessories;camcorder;hdtv;digitalcameras;television;....;\n


NexTag
Comparison Shopping for Products Mortgages Travel Schools More
Sign In
sony camera
NexTag Search
All Categories
related searches: sony digital camera, sony 7.2 digital camera, digital camera, more...
All Categories : Electronics : Digital Cameras : sony camera
See additional matches in Electronics Accessories, Electronics or all categories
shopzilla
the smarter way to shop
Your account
Bookmark this page!
9/11
Home
Clothes
Home & Garden
Computers
Halloween
Jewelry
Toys
Electronics
Sports
All Depts.
Shop for sony camera
in All Departments
Find it!
Search took 0.010 seconds.
Home > sony camera
Email this page
We found sony camera in the following departments
sony camera Digital Cameras
sony camera Digital Camera Accessories
sony camera Cell Phones
sony camera Memory Cards
sony camera Camera Accessories
See all matches for 'sony camera'
sony camera Batteries
sony camera Camcorder Accessories
sony camera Home Security
sony camera CD & DVD Drives
sony camera Miscellaneous
Sponsored Links
Sony Camera at Amazon.com
Low prices on sony camera. Qualified orders over $25 ship free
amazon.com Overall Rating:
Sony Camera Online
(Buy SONY™ Direct) Save Online at Sony's Online Store - Sony Camera
www.sonystyle.com Overall Rating:
Camera On Sale
Save 40-80% on Quality Film Cameras. Only $2.95 Shipping!
overstock.com Overall Rating:
Camera
Top Camera Offers Save on Camera
camera.pages.us.com Overall Rating: Not Yet Rated
Try eBay
Whatever you're looking for you can get it on eBay.
www.ebay.com Overall Rating:
Sony Camera

Simple query rewrite example

Semantic rule base

Assuming we have a brand and category dictionary, it is very easy to set up a rule base that does the necessary query rewrites:

```
# Shopping rules
@default
@automata(etc/vespa-rules.fsa)

[brand] -> brand:[brand];
[thisis] -> thisis:[thisis];

[brand] :- [C];
[thisis] :- [B];
```

http://qrs:5810?query=sony+camera&sorting&tracelevel=1&tracelevel.rules=1

```
− <meta type="trace">
  − <p>
      Transforming 'AND ;C:sony ;B:camera' to 'AND brand:sony ;B:camera' since '[brand] -> brand:[brand]' matched
    </p>
  − <p>
      Transforming 'AND brand:sony ;B:camera' to 'AND brand:sony thisis:camera' since '[category] -> thisis:[category]' matched
    </p>
  − <p>
      SemanticSearcher: Rewrote query: [AND brand:sony thisis:camera]
    </p>
  + <p></p>
```

Example: sony camera w/pricesort

Without rewrite rules (84 documents)

- Sony Ericsson Z520 Cingular, **pfrom = 10**
- Sony Ericsson Z520a (**Video Phone**) for Cingular w/ 2yr Contract- Free Shipping, **pfrom = 10**
- Sony NH AA 2DB –Camera **battery**- rechargeable- AA NiMH x 2- 2100 mAh, **pfrom = 6 350**
- 2 Pack NiMH AA Rechargeable **Battery**; For Various Digital Cameras OEM Equivalent to Sony, **pfrom = 10 950**
- Sony digital camera **battery**, CCD GR1, RUVI, DCR PC1, DCR PC3 and others. PN: NP F21 NPF21, **pfrom = 14 500**
- **Battery** for Sony DSC- T1 Digital Cameras models PN: NP FT1 T1, **pfrom = 19 000**
- Digital camera **battery** for Sony digital cameras **pfrom = 19 000**
- Digital Camera; Lithium Ion **battery** for Sony DCR series, **pfrom = 19 000**
- SanDisk **Memory Stick** Pro Duo 128MB with Adapter (MS Card 128MB), Compatible with Sony PSP and Sony Ericsson Cell Phone, **pfrom = 23 950**
- Digital camera **battery** for Sony Cybershot and other digital cameras, **pfrom = 24 950**.

2 phone contracts, 7 batteries 1 memory stick

None of the desired objects

With rewrite rules (13 documents)

1. Sony Waterproof Infrared Illumination **Security Camera** (480 TV Lines), **pfrom = 78 000**
2. CB25D 12V/24V Color **CCD Camera** 420TV Lines, **pfrom = 85 000**
3. Sony DSC P50 Cyber-shot 2MP **Digital Camera** with 3x Optical Zoom Silver (Part#: DSC P50), **pfrom = 134 000**
4. Sony DSC P52 Cyber-shot 3.2MP **Digital Camera**, **pfrom = 149 990**
5. CV 7911XH 1/3"Sony Ex View HAD CCD Day Night **Color Camera**, **pfrom = 179 000**
6. SONY CyberShot DSC W1, **pfrom = 199 950**
7. C3326EX Mini Professional 1/3" **Camera**; 12V DC, 270K Pixels SONY Chip Set, **pfrom = 235 000**
8. Sony DSC L1 Digital **Camera**, **pfrom = 259 000**
9. SONY DSC W7, **pfrom = 275 950**
10. SNC M3W Wireless Mini Pan/Tilt **Indoor Camera**, **pfrom = 355 540**

10 sony cameras sorted on price

Less recall, but dramatically improved relevancy

Evaluation: Query impact

Altered resultsets when using rules (1k queries/100hits)

- 71-75% document total setsize overlap
- Normal queries:
  - 65-70% resultset positions kept => **30-35% impact**
    - » Note: Large resultsets (relevancy measured on top 10%)
- Price-sort queries (prototype index estimate)
  - Still 74% setsize overlap
  - 52% resultpositions kept => **48% impact**

Evaluation: General relevancy impact

…not price sort queries…added bonus…

- Manual evaluation
  - Only judge queries understood: 60 random queries w/rules on/off (1200 judgements)
  - Scoring scheme:
    - 0 => 1 : Good to excellent documents
    - 0 : Can't say really
    - -1=> 0 : Awful to bad documents (remove)

Accumulated score: 240.5! vs 108.8

» 121% relative increase! ☺

TakeAway

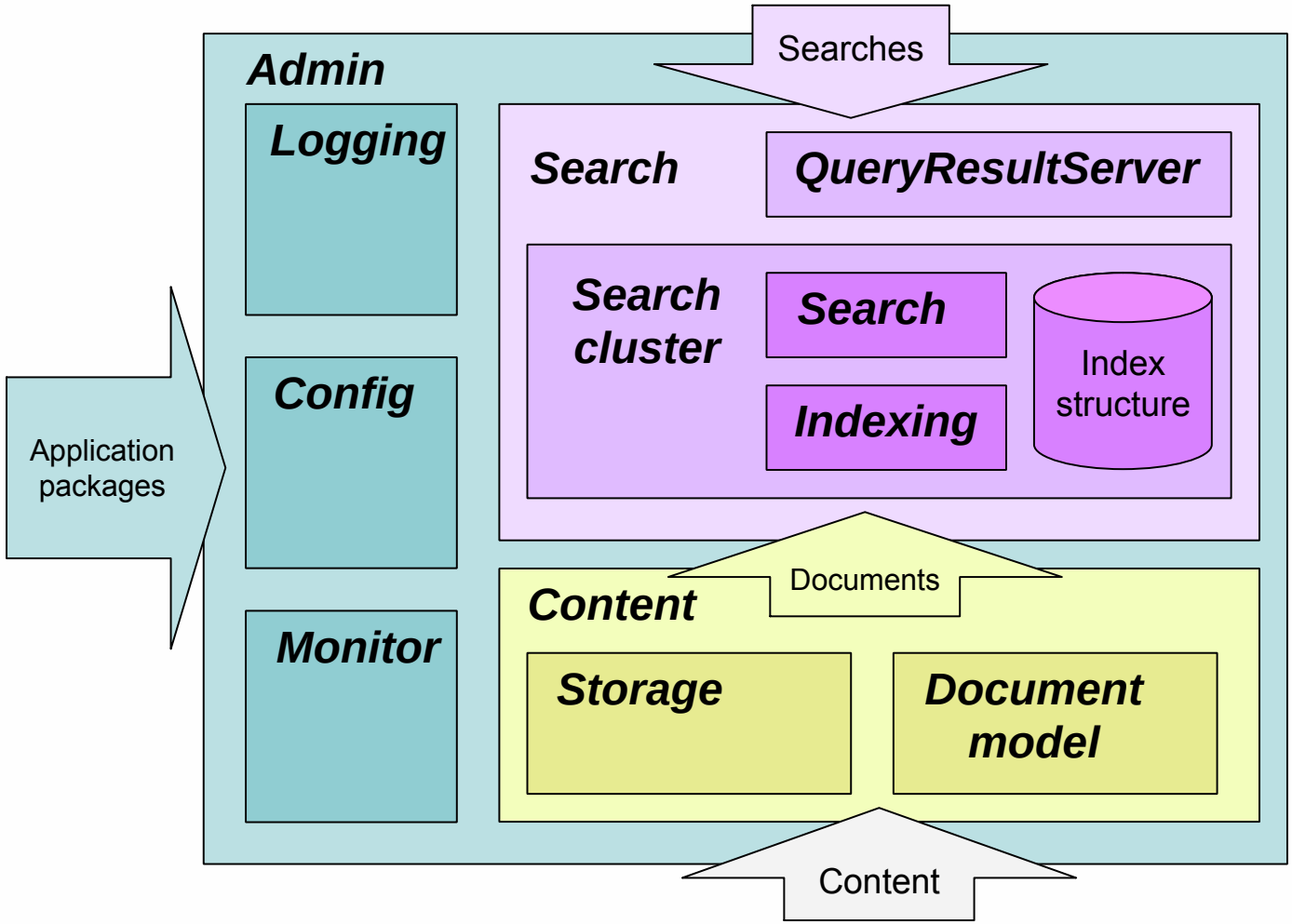
A few recognized query terms can have tremendous impact on relevance when they match indexed structured domain knowledge.

- ***i.e. semantic tags matching user language***

- … structured query rewrites scales well…
- .... but not understanding natural language yet ...
- … focus on semantic data mining & entity normalization for focused large scale search…


LIFE ENGINE™

Extra slide 1: Vespa platform components



Extra slide 2: Experimental framework for shopping

- **Prototype** system: single node
  - Paidmerchant data, many fields => limited data set
- **Evaluation** system: 52 nodes
  - 1 of 10 qrs (query result servers) had the query rewrites
  - Full, but old, shopping data set
  - Used for other test purposes
    - Q/A of 2.1.X (X = 1..6) for shopping
- Evaluation queries: Random samples of
  - QRS queries 500q/1000hits => overlap / similarity
  - Query units: 1000q/100 hits => semantic perfomance